Rheumatoid Arthritis Phenotype Algorithm

Harvard eMERGE

|  |  |
| --- | --- |
| Document Version: | 1.0.00-001 |
| RA\_filterpositive\_19Jul2017 | 19Jul2017 |

Table of Contents

[Section 1. Overview 3](#_Toc495063049)

[1.1. Performance Characteristics 3](#_Toc495063050)

[Section 2. Algorithm Definition 4](#_Toc495063051)

[2.1. Feature Dictionary 4](#_Toc495063052)

[2.2. Model Formula 5](#_Toc495063053)

[2.3. Defining RA Cases 6](#_Toc495063060)

[2.4. Defining RA Controls 7](#_Toc495063061)

[2.5. References 7](#_Toc495063062)

[Section 3. Appendices 8](#_Toc495063063)

[3.1. Phenotype Workflow 8](#_Toc495063064)

[3.2. Feature Definitions 8](#_Toc495063065)

[3.3. Control Exclusion Codes 10](#_Toc495063071)

[3.4. RA Feature Distribution Data Dictionary 11](#_Toc495063072)

[3.5. RA Demographics Data Dictionary 12](#_Toc495063073)

# Overview

This rheumatoid arthritis (RA) algorithm was created using a machine-learning logistic regression model.

First, the overall data set was screened to identify subjects with at least one RA diagnosis code. Chart reviews were performed among the screen-positive set by a clinical expert to establish a gold standard, (based on the 2010 American College of Rheumatology (ACR) criteria for classification of rheumatoid arthritis1.)

The chart review results were used to train the regression model.  A set of electronic health record (EHR) attributes, (also called features), associated with the disease, were defined using Automatic Feature Extraction (AFEP)2. A penalized logistic regression model (LASSO) was trained to classify RA in the gold-standard Training Set (3,4). The regression model identified the relative weights (beta coefficients) of the features significantly associated with RA. Once the model was created, a threshold value based on specificity = 97% and positive predictive value (PPV) was selected for to identify cases and controls, with those above the cutoff identified as cases and those below, non-cases or controls.  The PPV is the rate of true positives in those classified as cases by the algorithm.

An independent Test Set was created by chart reviews of a randomly select set of 100 screen positive charts. In addition, a secondary validation was performed in an independent Test Set from Vanderbilt that included 290 charts reviewed for ACR criteria for rheumatoid arthritis (154 Definite RA, 136 Not RA)5. For a detailed description, please refer to the Harvard eMERGE Algorithm Training and Testing document.

This document describes how to implement the rheumatoid arthritis (RA) phenotype algorithm created at Harvard. The model predicts current and past history of disease with no temporal constraints.

## Performance Characteristics

The overall model AUC is 0.95

For implementation of the algorithm, specificity is set at 0.97 (FPR=0.03), where PPV = 0.946 and NPV = 0.923 in the training set.

#### RA Algorithm Performance Characteristics

|  |  |  |
| --- | --- | --- |
| PPV (precision) | Positive Predictive Value (PPV) | 0.946 |
| TPR (sensitivity) | Sensitivity or True Positive Rate (TPR) | 0.866 |
| FPR (1-specificity) | False Positive Rate (FPR) | 0.03 |
| NPV | Negative Predictive Value (NPV) | 0.923 |

# Algorithm Definition

The algorithm definition consists of 3 parts:

1. Feature Dictionary
2. Model Formula
3. Defining RA Cases

## Feature Dictionary

The table below lists the features for the RA phenotype algorithm and their associated beta coefficients (weights). The weights are used to derive a predicted probability of RA or no RA for each subject. The list of codes for each feature is listed in the Appendices (Section 3).

Feature\_ID Nomenclature: [Phenotype] + ‘\_COD\_’ + [Feature Type Abbreviation] , where ‘COD’ means coded variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature\_ID | Short\_Name | Transformation | Weight  (Beta) | Feature Description |
| RA\_COD\_DX\_RheumatoidArthritis\_v2 | DX\_RA | log1p | 1.937 | Coded mentions of a Rheumatoid Arthritis diagnosis per subject |
| RA\_COD\_DX Systemiclupuserythematosus | DX\_Lupus | log1p | -0.529 | Coded mentions of a Lupus diagnosis per subject |
| RA\_COD\_DX\_PsoriaticArthritis\_v2 | DX\_Psoriatic | log1p | -0.122 | Coded mentions of a Psoriatic arthritis diagnosis per subject |
| RA\_COD\_LAB\_RFpos1 | LAB\_RF | n/a | 1.639 | A lab test for Rheumatoid Factor.  1 = subject has had the text and it is positive.  0 = no test or RF negative |
| patient\_dxenct | patient\_dxenct | log1p | -0.954 | Total number of encounters (visits), per subject, with a coded diagnosis (any diagnosis not limited to RA). |

## Model Formula

The model formula is used to calculate the predicted probability for each subject using the list of features and weights (betas). The subject’s feature scores are summed to get the subject’s overall score, which is converted into a predictive value.

### Steps to calculate the predicted probability

### Create a summary table with feature counts by subject

For each feature, except Lab\_RF, count the number of distinct dates for which each subject has a code. For Lab\_RF, 1 = RF positive and 0 = RF negative or no test.

**Example:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SUBJID** | **Patient\_dxenct** | **DX\_RA** | **DX\_Lupus** | **DX\_Psoriatic** | **Lab\_RF** |
| 123 | 208 | 35 | 0 | 0 | 1 |
| 124 | 474 | 18 | 0 | 0 | 1 |
| 125 | 298 | 1 | 33 | 0 | 1 |
| 126 | 269 | 180 | 2 | 0 | 0 |
| 127 | 152 | 51 | 0 | 0 | 0 |
| 128 | 115 | 0 | 0 | 0 | 0 |

### Flag subjects that meet filter criteria

To be included in the algorithm, subjects must have a count of one or more RA diagnoses (Section 3.2.1).

**Example:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SUBJID** | **Patient\_dxenct** | **DX\_RA** | **DX\_Lupus** | **DX\_Psoriatic** | **LAB\_RF** | **Meets Filter Criteria** |
| 123 | 208 | 35 | 0 | 0 | 1 | Y |
| 124 | 474 | 18 | 0 | 0 | 1 | Y |
| 125 | 298 | 1 | 33 | 0 | 1 | Y |
| 126 | 269 | 180 | 2 | 0 | 0 | Y |
| 127 | 152 | 51 | 0 | 0 | 0 | Y |
| 128 | 115 | 0 | 0 | 0 | 0 | N |

### Calculate subjects’ feature scores (beta values).

The **Transformation** and **Weight** for each feature can be found in the feature dictionary (Section 2.1).

The **Patient\_Ct** is taken from the summary table (Section 2.2.1.1) for each feature.

Formula for Beta Value:

IF Transformation = ‘Log1P’ THEN **Log(patient\_ct + 1) \* weight**

ELSE **patient\_ct** (or value) **\* weight**

*(Log() corresponds to the natural logarithm)*

#### Example for subject 123:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SUBJID** | **Feature** | **Transformation** | **Patient\_Ct** | **Weight** | **Applied Formula** | **Beta Value** |
| 123 | **RA\_COD\_DX\_RheumatoidArthritis\_v2** | Log1p | 35 | 1.937 | Log(35 + 1) \* 1.937 | 6.9412 |
| 123 | **RA\_COD\_DX\_Systemiclupuserythematosus** | Log1p | 0 | -0.529 | Log(0 + 1) \* -0.529 | 0 |
| 123 | **RA\_COD\_ DX\_Psoriaticarthritis\_v2** | Log1p | 0 | -0.122 | Log(0 + 1) \* -0.122 | 0 |
| 123 | **RA\_COD\_LAB\_RFpos1** | n/a | 1 (bit 0 or 1) | 1.639 | 1 \* 1.639 | 1.639 |
| 123 | **patient\_dxenct** | Log1p | 208 | -0.954 | Log(208 + 1) \* -0.954 | -5.0965 |

### Calculate the beta values per subject

Sum the feature beta values by subject.

Example for Subject 123:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SUBJID** | **Feature** | **Transformation** | **Patient\_Ct** | **Weight** | **Applied Formula** | **Beta Value** |
| 123 | **RA\_COD\_DX\_RheumatoidArthritis\_v2** | Log1p | 35 | 1.937 | Log(35 + 1) \* 1.937 | 6.9412 |
| 123 | **RA\_COD\_DX\_Systemiclupuserythematosus** | Log1p | 0 | -0.529 | Log(0 + 1) \* -0.529 | 0 |
| 123 | **RA\_COD\_ DX\_Psoriaticarthritis\_v2** | Log1p | 0 | -0.122 | Log(0 + 1) \* -0.122 | 0 |
| 123 | **RA\_COD\_LAB\_RFpos1** | n/a | 1 (bit 0 or 1) | 1.639 | 1 \* 1.639 | 1.639 |
| 123 | **patient\_dxenct** | Log1p | 208 | -0.954 | Log(208 + 1) \* -0.954 | -5.0965 |
|  |  |  |  |  | **Sum Beta Value**  *(Does not include Intercept Beta\_Value*) | **3.4837** |

### Calculate subjects’ Predicted Probability

The beta value for the intercept is not calculated, it is an assigned weight. The RA Intercept = **-1.017**

Predicted Probability = **exp(Intercept + sum\_beta)/ (1 + exp(Intercept + sum\_beta))**

Example for Subject 123:

exp(**-1.017** + **3.4837**)/ (1 + exp(**-1.017** + **3.4837**)) =  **0.92177414**

Subject 123’s predicted probability = 0.9217

*(exp() = exponential value)*

## Defining RA Cases

Once each patient has a predicted probability score, cutoffs are applied to classify a patient as having the phenotype. The cutoffs correspond to a specificity or PPV level estimated during the model training.

Categorize the subjects as an RA case based on the cutoff.

Rheumatoid Arthritis Cutoffs:

The subject is a **case** if their predicted probability is >= 0.632.

If their predicted probability is < 0.632, the subject is a non-case (NA).

#### Cutoff Examples:

|  |  |  |  |
| --- | --- | --- | --- |
| **SubjID** | **Predicted Probability** | **Cutoff** | **Case** |
| 123 | 0.9217 | >= 0.632 | Yes |
| 124 | 0.6095 | >= 0.632 | No |

## Defining RA Controls

Any subject that:

* DOES NOT have a code listed in the **RA\_COD\_DX\_RheumatoidArthritis\_v2** feature (Section 3.2.1)

AND

* DOES NOT have any code in the Controls Exclusion Codes table (Section 3.3)

is a **control**.

## References

1. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, Birnbaum NS, Burmester GR, Bykerk VP, Cohen MD, Combe B, Costenbader KH, Dougados M, Emery P, Ferraccioli G, Hazes JM, Hobbs K, Huizinga TW, Kavanaugh A, Kay J, Kvien TK, Laing T, Mease P, Ménard HA, Moreland LW, Naden RL, Pincus T, Smolen JS, Stanislawska-Biernat E, Symmons D, Tak PP, Upchurch KS, Vencovský J, Wolfe F, Hawker G. 2010 Rheumatoid arthritis classification criteria: an American College Rheumatology/European League Against Rheumatism collaborative initiative.

Arthritis Rheum. 2010 Sep;62(9):2569-81. doi: 10.1002/art.27584. PubMed PMID: 20872595.

2. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc.* Sep 2015;22(5):993-1000.

3.Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association.* 2014/10/15 2006;101(476):1418-1429.

4. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika.* 2007;94(3):691-703.

5. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H, Karlson EW, Perez RG, Gainer VS, Murphy SN, Ruderman EM, Pope RM, Plenge RM, Kho AN, Liao KP, Denny JC. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012 Jun;19(e1):e162-9. Epub 2012 Feb 28. PubMed PMID: 22374935; PubMed Central PMCID: PMC3392871

# Appendices

## Phenotype Workflow



## Feature Definitions

### RA\_COD\_DX\_RheumatoidArthritis\_v2 (Number of coded Rheumatoid Arthritis diagnoses)

|  |  |  |
| --- | --- | --- |
| Feature Type | Code | Name |
| Diagnosis (ICD9) | 714 | Rheumatoid arthritis and other inflammatory polyarthropathies |
|  | 714.0 | Rheumatoid arthritis |
|  | 714.1 | Felty's syndrome |
|  | 714.2 | Other rheumatoid arthritis with visceral or systemic involvement |
| Diagnosis (ICD10) | M05\* | Rheumatoid arthritis with rheumatoid factor *(include all codes starting with M05)* |
|  | M06\* | Other rheumatoid arthritis *(include all codes starting with M06)* |

### RA\_COD\_DX\_SystemicLupusErythematosus (Number of coded Lupus diagnoses)

|  |  |  |
| --- | --- | --- |
| Feature Type | Code | Name |
| Diagnosis (ICD9) | 710 | Diffuse diseases of connective tissue |
|  | 710.0 | Systemic lupus erythematosus |
| Diagnosis (ICD10) | M32.1\* | Systemic lupus *(include all codes starting with* M32.1) |

### RA\_COD\_DX\_PsoriaticArthritis\_v2 (Number of coded Psoriatic Arthritis diagnoses)

|  |  |  |
| --- | --- | --- |
| Feature Type | Code | Name |
| Diagnosis (ICD9) | 696.0 | Psoriasis and similar disorders |
| Diagnosis (ICD10) | L40.5\* | Psoriasis *(include all codes starting with* L40.5) |

### RA\_COD\_LAB\_RFpos1

|  |  |  |
| --- | --- | --- |
| Feature Type | Code | Name |
| LOINC | 15205-8 | Rheumatoid factor [Units/​volume] in Serum by Nephelometry Rheumatoid factor [Units/Volume] in Serum by NephelometryRheumatoid factor [Units/​volume] in Serum by Nephelometry |

### Patient\_dxenct

The total number of encounters with a coded diagnosis.

* Includes all diagnosis codes
* NOT limited to the diagnosis codes used within the phenotype (all phenotypes will have the same value for Patient\_dxenct)
* Limit to one occurrence per date
* An inpatient stay that spans multiple days will be counted as one date

## Control Exclusion Codes

|  |  |  |
| --- | --- | --- |
| Name | ICD9 | ICD10 |
| Ankylosing spondylitis | 720.0, 720.2, 720.8, 720.81, 720.89, 720.9 | M45, M45.0, M45.1, M45.2, M45.3, M45.4, M45.5, M45.6, M45.7, M45.8, M45.9, M46.1, M46.90, M08.1 |
| Polymyalgia rheumatica | 725, 725.0 | M31.5, M35.3 |
| Rheumatoid arthritis (RA) | 714.0, 714.8, 714.89, 714.9, 714.2, 714.81, 714.3 | M05\*, M06\* |
| JIA | 714.30, 714.31, 714.32, 714.33 | M08.00, M08.2, M03.3, M08.40 M08.9, M08.8 |
| Felty's syndrome | 714.1 | M05.00 |
| Rheumatic fever | 390, 391, 392, 392.9, 390.9, 714.4 | i00, M12.0, M12.00, M12.01, M12.011, M12.012, M12.019, M12.02, M12.021, M12.022, M12.029, M12.03, M12.031, M12.032, M12.039, M12.04, M12.041, M12.042, M12.049, M12.05, M12.051, M12.052, M12.059, M12.06, M12.061, M12.062, M12.069, M12.07, M12.071, M12.072, M12.079, M12.08, M12.09 |
| Palindromic rheumatism (PR) | 719.3, 719.30, 719.31, 719.32, 719.33, 719.34, 719.35, 719.36, 719.37, 719.38, 719.39 | M12.3, M12.37, M12.32, M12.34, M12.35, M12.36, M12.372, M12.322, M12.342, M12.352, M12.362, M12.312, M12.322, M12.39, M12.38, M12.371, M12.321, M12.341, M12.351, M12.361, M12.311, M12.331, M12.31, M12.379, M12.329, M12.349, M12.359, M12.369, M12.319, M12.30, M12.339, M12.33 |
| Psoriatic arthritis | 696.0 | L40.54, M07.0, M07.3 |
| Lupus erythematosus, systemic | 710.0, 695.4 | M32.14, M32.13  L93, M32.8, L93.2, M32.19, M32.12, M32.10, M32, M32.9  M32.1, M32.1, M35.1 |
| Unspecified polyarthropathy or polyarthritis | 716.50, 716.51, 716.52, 716.53, 716.54, 716.55, 716.56, 716.57, 716.58 | M06.4, M13.0 |
| Synovitis | 727.0, 727.00, 727.01 | M65.8\*, M65.9 |

## RA Feature Distribution Data Dictionary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **VARNAME** | **VARDESC** | **TYPE** | **REQUIRED** | **VALUE** |
| SUBJID | The eMERGE unique ID of Subject | String | Yes |  |
| Case\_Control | Whether the subject is a Case or Control | string | Yes | C49152 = Case; C28143= Control  NA = Non-case  CX=Control Excluded |
| Patient\_dxent | Total number of encounters (visits), per subject, with a coded diagnosis (any diagnosis, not limited to RA). | Integer | Yes |  |
| RA\_COD\_DX\_RheumatoidArthritis\_v2\_Count | The Rheumatoid Arthritis feature count of distinct dates in which a subject has a code from this feature. | Integer | Yes |  |
| RA\_COD\_DX\_Systemiclupuserythematosus\_Count | The Lupus count of distinct dates in which a subject has a code from this feature. | Integer | Yes |  |
| RA\_COD\_DX\_Psoriaticarthritis\_v2\_Count | The Psoriatic Arthritis feature count of distinct dates in which a subject has a code from this feature. | Integer | Yes |  |
| RA\_COD\_LAB\_RFpos1 | The Rheumatoid Factor (RF) lab feature flag (1=positive, 0=negative or missing). | Integer | Yes |  |
| Predicted\_Probability | The predicted probability score for the subject. | Decimal | Yes |  |

## RA Demographics Data Dictionary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **VARNAME** | **VARDESC** | **TYPE** | **REQUIRED** | **VALUE** |
| SUBJID | Deidentified subject/eMERGE ID | String | Yes |  |
| Case\_Control | Case Control Status | string | Yes | C49152=Case;  C28143=Control;  Non-case;  Control Excluded |
| Sex | Sex of the participant | String | Yes | C46119=Male; C46110=Female; U=Unknown; NA=Not Assessed; .=Missing |
| Race | Race of the participant | String | Yes | C16352=Black or African American; C41259=American Indian or Alaska Native; C41260=Asian; C41261=White; C41219=Native Hawaiian or other Pacific Islander; C17998=Unknown; C43234=Not Reported |
| Ethnicity | Ethnicity of the participant | String | Yes | C17459=Hispanic or Latino; C41222=Not Hispanic or Latino; C41221=Unknown; .=Missing |
| Birth\_year | Four digit year of birth | integer | Yes |  |